An overview of intelligent data assistants for data analysis

Serban, F; Kietz, J U; Bernstein, A

Abstract: Today's intelligent data assistants (IDA) for data analysis are focusing on how to do effective and intelligent data analysis. However this is not a trivial task since one must take into consideration all the influencing factors: on one hand data analysis in general and on the other hand the communication and interaction with data analysts. The basic approach of building an IDA, where data analysis is (1) better as well as (2) faster in the same time, is not a very rewarding criteria and does not help in designing good IDAs. Therefore this paper tries to (a) discover constructive criteria that allow us to compare existing systems and help design better IDAs and (b) review all previous IDAs based on these criteria to find out what are the problems that IDAs should solve as well as which method works best for which problem. In conclusion we try to learn from previous experiences what features should be incorporated into a new IDA that would solve the problems of today's analysts.

An overview of intelligent data assistants for data analysis

Serban, F; Kietz, J U; Bernstein, A

An overview of intelligent data assistants for data analysis

Abstract

Today's intelligent data assistants (IDA) for data analysis are focusing on how to do effective and intelligent data analysis. However this is not a trivial task since one must take into consideration all the influencing factors: on one hand data analysis in general and on the other hand the communication and interaction with data analysts. The basic approach of building an IDA, where data analysis is (1) better as well as (2) faster in the same time, is not a very rewarding criteria and does not help in designing good IDAs. Therefore this paper tries to (a) discover constructive criteria that allow us to compare existing systems and help design better IDAs and (b) review all previous IDAs based on these criteria to find out what are the problems that IDAs should solve as well as which method works best for which problem. In conclusion we try to learn from previous experiences what features should be incorporated into a new IDA that would solve the problems of today's analysts.

An Overview of Intelligent Data Assistants for Data Analysis

Floarea Serban and **Jörg-Uwe Kietz** and **Abraham Bernstein**[1]

Abstract. Today's intelligent data assistants (IDA) for data analysis are focusing on how to do effective and intelligent data analysis. However this is not a trivial task since one must take into consideration all the influencing factors: on one hand data analysis in general and on the other hand the communication and interaction with data analysts. The basic approach of building an IDA, where data analysis is (1) better as well as (2) faster in the same time, is not a very rewarding criteria and does not help in designing good IDAs. Therefore this paper tries to (a) discover constructive criteria that allow us to compare existing systems and help design better IDAs and (b) review all previous IDAs based on these criteria to find out what are the problems that IDAs should solve as well as which method works best for which problem. In conclusion we try to learn from previous experiences what features should be incorporated into a new IDA that would solve the problems of today's analysts.

1 Introduction

As technology advances generating larger amounts of data becomes easier, increasing the complexity of data analysis. The process of analyzing data is not a trivial task requiring time, experience and knowledge about the existing operators. As new types of data appear and corresponding algorithms are designed to analyze them, the task of the data analyst becomes more complex since it is requiring awareness of a continuos expanding set of operators [32].

Several data analysis (DA) systems have been developed to analyze data, each of them providing a large number of operators (Clementine, SAS Enterprise Miner, RapidMiner, Weka). Even if they are constantly trying to improve the quality of the analysis, the number of operators and the data size are growing making it difficult for users to find an appropriate sequence of operators for a given task or data set. Therefore with the continuous evolution of these systems there is an essential need to provide more assistance to the users and suggest the right operators for their current task. Even if this need has already been mentioned in the literature [22], to our knowledge, there is no IDA which could successfuly solve this problem. Several attempts in developing IDAs have been made [52, 4, 21] but they were either just a proof of concept [4, 20] or they are no longer maintained [52]. Thus they are unusable for today's data analysts. Moreover they are focusing either on novice analysts or on experts and therefore cannot provide support for all users. Furthermore they are either guiding the user step by step or leaving her free to make her own decisions, so the user is either fully restricted to a fixed set of steps or left alone in a jungle of operators. Therefore finding the right sequence of operators becomes almost impossible.

It is difficult to precisely define how the best IDA should be designed. Nevertheless we can say that the basic (minimal) features of an IDA should be defined in terms of quality and the amount of time spent for the analysis . Hence an IDA is basically defined by its goals: (1) to analyze the data better with the IDA than without, given an amount of time, and (2) to do it faster with the IDA support than without, given a required level of quality. The quality of the analysis is important since an IDA should improve the quality and help the user obtain better results. In addition time is always a problem, large data sets take a long of time to be analyzed therefore the assistant should decrease this time and do the analysis faster. An IDA should definitely include these two features, but they are not sufficient to design a good system. Therefore in this paper we develop a set of constructive criteria that allow us to (a) compare existing systems and (b) help design a better IDA. Moreover we review all the previous attempts to develop IDAs based on these criteria to find out (a) which problems IDAs should solve and (b) which method works best for which problem. Furthermore, based on the comparison with existing IDAs, we present a set of features for a new IDA, called EG-IDA (explorative guidance IDA) which helps the user explore and effectively use the data analysis techniques.

The rest of the paper is organized as follows: Section 2 presents existing work on IDAs or intelligent systems developed for data analysis, makes an overview of their features and their applicability in current scenarios and highlights some of their limitations, then Section 3 introduces the desired features of a new IDA that should overcome the limitations of existing IDAs and finally conclusions and future work are described in Section 4.

2 IDAs evolution

Over the course of time scientists worked on developing systems to improve data analysis. One of the first research directions was to automate the data analysis process as well as provide more user support, therefore introducing the Statistical Expert Systems (SES) at the beginning of '80s [33, 25]. At that time the data analysis process was mainly based on statistics; statisticians were studying the relationships between variables. Systems like REX [24] tried to improve the user support for data analysis by using the knowledge from expert users and generating rules. These rules were used to help users in solving data analysis problems like linear regression. The advice was limited to the encoded expert knowledge and could be applied only on a reduced set of problems. Moreover the systems could not handle more complex questions and were restrained to a hardcoded set of answers.

[1] University of Zurich, Department of Informatics, Dynamic and Distributed Information Systems Group, Binzmühlestrasse 14, CH-8050 Zurich, Switzerland {serban|kietz|bernstein}@ifi.uzh.ch

This led to an extension of statistical expert systems, the knowledge enhancement systems (KES) which try to offer a more flexible access to statistical expertise. Such an approach is that of KENS [35], a KES for non-parametric statistics, which assists the user in solving problems and tries to improve her understanding of nonparametric statistics. The user can ask questions in natural language and the system provides a list of answers based on both human and machine expertise. The successor of KENS is NONPAREIL [35] which still focuses on nonparametric statistics but it uses hypertext links instead of searching. A similar approach is LMG [35] which assists the user in fitting linear models. In fact LMG presents to the user questions which contain hypertext buttons through which she can have access to explanations about a selected concept. The KES systems give more freedom to the users in exploring the statistical world. They also represent learning environments - the user can easily learn by questioning the system how to handle different problems from the domain (either nonparametric statistics or linear modeling).

The next evolution step was the appearance of Intelligent Exploratory Data Analysis systems. Systems like AIDE [52] or Data Desk [53] offer intelligent support for exploratory data analysis. They provide means that make the analyst's task easier by improving the user interactivity or the user support with more explanations. Finally the evolution continued with more focus on the IDAs in systems like IDEA [4], CITRUS [54], MetaL [41]. Remarkable work was done by Robert Engels who describes in his PhD thesis the User Guidance Module (UGM) [23] - a cookbook on how to do Data Mining as well as a written and also implemented (part of CITRUS) IDA. Another "written" IDA is the CRISP-DM process model [13] since it is a step-by-step data mining guide - it is considered the standard process model for Data Mining. Even if CRISP-DM is just a standard by the fact that it presents guidance on how to do data mining it constitutes itself as an IDA.

The present survey covers several different systems[2] as shown in Table 1 that can be grouped in the following categories: Statistical Expert Systems (SES), Knowledge Enhancement Systems (KES), Exploratory Data Analysis Systems (EDA), Data Analysis Systems (DAS) and Automatic Data Assistants (ADA). The comparison was done based on the published research papers for the SES and EDA since the systems are no longer available. For DA systems we were able to try and check the functionality of the current systems, Clementine, RapidMiner, Weka, SPSS, Excel, Matlab and R which facilitate access to a collection of algorithms but they offer no real decision support to inexperienced end-users.

2.1 Comparison of existing IDAs

All the enumerated systems do intelligent data analysis since they offer guidance (partially) and help the user to analyze her data. We identified seven possible dimensions on which the systems could be compared:

- Support for modeling a **single-step** from the KDD process vs. **multi-step** KDD process
  We compare systems which help the user to model a single step from the KDD process by guiding her on how to use a specific operator, how to choose the right parameters for it, to systems which provide support for multi-step processes - they assist the user in the selection and application of available techniques at each step of the DM process. But an IDA should include both these dimensions: the user needs information on configuring a specific operation as well as building the sequence of steps from the KDD process.
- **Graphical editing** vs. **automatic generation**
  Graphical editing refers to enabling the user to draw the process manually - choose the operators, set the right inputs/outputs and parameters. While automatic generation provides the user with a set of workflows (or at least one) that she needs to execute in order to solve the data mining task. The system automatically sets all the inputs/outputs and the parameters for each operator. This is useful for users that don't have too much experience with the data mining operators. Based on the data and a description of their task they get a set of possible scenarios of solving a problem. Both of the criteria are recommended for an IDA since the users have different experience and knowledge and they need different help.
- **Re-use past experiences** vs. **generation from scratch**
  The system saves all the produced cases and records the ones which have succeeded or have failed to solve the given task. Reusing past cases improves the generation of new better workflows since the system reuses only the similar cases or parts of the cases that were successful. Moreover it saves time because the system doesn't have to generate each workflow from scratch and also helps avoiding the repetition of mistakes. Reusing past cases is definitely an asset for an IDA since it can improve recommendations and should be considered when implementing such an IDA. An IDA should definitely include the first feature, being able to reuse cases can improve the assistants recommendations.
- **Task decomposition** vs. **plain plan**
  Task decomposition structures and breaks down the complexity of a KDD task. It originates from the field of knowledge acquisition where it was used to describe and specify complex tasks [21]. Task descriptions can be reused thus decreasing the development time and simplifying the process of decomposing a KDD task. Also the task role is to transform the initial problem with certain features into the goal problem with additional features [12].
- **Design support** vs. **explanations for result/output**
  By design support we refer to the help and advice the system provides during the design of the data mining process. The user can easily find information about operators, he is given input or hints on how to solve problems or errors. Opposed to design support the explanations help the user interpret the results by suggesting different methods (e.g., graphs). Explanation includes the support for the interpretation of intermediate and final results as well as the capability to explain the reasoning and decisions of the system. Both features should be included in IDAs since they help the user exploring and discovering the operators.
- **Experimental** vs. **analytical** approach
  Experimental systems enable the user to execute the workflow, she or the system creates, and visualize the results. Contrary to experimental systems analytical ones are based on rules learned either from experts or from data characteristics. But both are required for an IDA - one enables the user to execute operators and the other one can give recommendations on how and when to use the operators.
- **Data mining experience level**
  The systems can be used by users with a certain DM knowledge level; we have systems which consider naive users, others which can be used only by experienced users, by experts or by users with any level of knowledge. Also for some systems the intended user is not specified or the systemy can't be used since they are just a proof of concept.

[2] For systems without have name we considered the name of the main author.

Category	System name	KDD single step support	KDD multi-step support	Graphical workflow editing	Automatic workflow generation	Re-use past experiences	Task decomposition	Design support	Explanations	Experimental	Analytical	DM experience level	References
SES	REX	++	−	−−	−−	−−	−−	−−	++	−	++	naive	[24]
	SPRINGEX	++	−	−−	−−	−−	−−	−−	+	−	++	experienced	[49]
	STATISTICAL NAVIGATOR	++	−	−−	−−	−−	−−	−−	++	−	++	experienced	[49]
	MLT Consultant	++	−	−−	−−	−−	−−	−−	++	++	++	all	[50]
KES	KENS	++	−	−−	−−	−−	−−	−−	+	−	++	experienced	[34, 35]
	LMG	++	−	−−	−−	−−	−−	−−	+	−	++	all	
	GLIMPSE	++	−	−−	−−	−−	−−	−−	++	−	++	experienced	[55, 56]
EDAS	AIDE	+	+	−	+	+	+	−	+	++	−	experienced	[52]
	Data Desk	0	0	−	−	−	−	−	−	++	−	all	[53]
DAS	SPSS	0	0	−	−	−	−	−	+	++	−	experienced	
	Clementine	0	0	++	−	−	−	++	++	++	−	experienced	
	SAS Enterprise Miner	0	0	++	−	−	−	++	++	++	−	experienced	[11]
	Weka	0	0	++	−	−	−	+	−	++	−	experienced	[30]
	RapidMiner 5.0	0	0	++	−	−	−	++	+	++	−	experienced	[46]
	KNIME	0	0	++	−	−	−	++	+	++	−	experienced	[5]
	Orange	0	0	++	−	−	−	++	+	++	−	experienced	[17]
	R	0	0	−	−	−	−	−	−	++	−	experienced	[40]
	MATLAB	0	0	−	−	−	−	+	−	++	−	experienced	[44]
	Excel	0	0	−	−	−	−	−	−	++	−	all	[43]
	Data Plot	0	0	−	−	−	−	++	−	++	−	unspecified	[37]
	GESCONDA	0	0	−	−	−	−	+	−	++	−	unspecified	[28, 27]
ADAS	IDEA	−	++	−	++	−	−	−	−	++	−	unspecified	[4]
	CITRUS	−	++	−	++	++	++	−	++	++	−	all	[21, 22]
	Záková	−	++	−	++	−	−	−	−	++	−	unspecified	[58, 57]
	KDDVM	−	++	−	++	−	−	−	−	+	−	experienced	[18, 19]
CBRS	METAL (DMA)	++	−	−	−	++	−	−	−	−	++	experienced	[41, 29]
	Mining Mart	0	0	+	−	++	−	−	−	++	−	experienced	[48, 42]
	Charest	0	0	−	−	++	−	−	−	++	−	all	[14, 15]
Other	CRISP-DM	+	++	−−	−−	−−	++	−−	−	−−	++		[13]
	IDM	++	−	−	−	−	−	++	+	+	+	unspecified	[6]
	MULREG	++	−	−	−	−	−	−	+	++	−	all	[20]

++ = well supported (a main feature of the system)
\+ = supported
0 = neutral, the system can do it but there is no assistance
− = not present but integrable
−− = not possible

Table 1: List of IDAs by category

2.1.1 Modeling a single-step from the KDD process vs. multi-step

Initially the systems provided support for modeling a single step from the KDD process, like SES and KES do. REX focuses on linear regression advice as opposed to the commercially available SES which have a broader application domain and can provide advice on several problems. Hence Springex handles bivariate, multivariate and non-parametric statistics and Statistical Navigator covers even more statistical techniques: multivariate causal analysis, scaling and classification, exploratory data analysis, etc.. The system gives advice based on the information provided by the user combined with the rules from the knowledge base. KES are more flexible from the point of view on how they present the advice - the user is free to question the systems and presents all the answers to the user. GLIMPSE even suspends when information is missing, and suggests actions to find that missing information. A more complex approach is that of MLT-Consultant [50] which provides advice on how to use a specific algorithm from the Machine Learning Toolbox (MLT). The advice is based on a knowledge base containing rules extracted from real-world tasks achieved by the domain experts and also from the interaction with the ML algorithm developers. As contribution the Consultant-2 provides support for preprocessing data as well as suggesting new methods after the one applied produced unsatisfactory results. Thus a step from the KDD process is seen not as a single-step but as a cyclic process where the user can reapply other algorithms if she is not satisfied with the current results. Moreover it is one of the first attempts to use meta-learning - they tried to suggest the appropriate ML tool based on the task description, the data characteristics and the output format. A similar view is the one of MULREG which recommends certain techniques for linear regression basing its advice on metadata (the measurement level) and on properties of the data themselves (parameters of their distribution). Also MetaL uses the notion of *meta-learning* to advise users which induction algorithm to choose for a particular data-mining task [38]. One of the outcomes of the project was a Data Mining Advisor (DMA) [29] based on meta-learning that gives users support with model selection. IDM has a knowledge module which contains meta-knowledge about the data mining methods, and it is used to determine which algorithm should be executed for a current problem. GESCONDA provides several tools for recommending a proper way to face the analysis in order to extract more useful knowledge like method suggestion, parameter setting, attribute/variable meta-knowledge management, etc. Other work was done by the StatLog project [45] which has investigated which induction algorithms to use given particular circumstances. This approach is further explored by [8, 26] which use meta-rules drawn from experimental studies, to help predict the applicability of different algorithms; the rules consider measurable characteristics of the data (e.g., number of examples, number of attributes, number of classes, kurtosis, etc.). [10] present a framework which generates a ranking of classification algorithms based on instance based learning and meta-learning on accuracy and time results.

DAS are neutral from this point of view - the user is free to choose an operator, set the parameters she wants to use and execute it. A totally opposite direction is the one of IDEA system which provides users with systematic enumerations of valid DM processes and rankings by different criteria. The enumeration is done based on the characteristics of the input data and of the desired mining result as well as on an operator ontology which specifies preconditions and effects for each operator. However the system doesn't in fact support the user through the steps of the DM process it just enumerates the steps.

The shift from single-step to multi-step started with the introduction of the CRISP-DM standard and the CITRUS system which helps the user through all the phases of the KDD process. Current DAS enable the user to design and execute multi-step KDD processes, but this becomes hard when the processes have a large number of operators.

2.1.2 Graphical editing vs. automatic generation

SPSS Clementine (SPSS Modeler nowadays), SAS Enterprise Miner [3] and RapidMiner 5.0 [4] enable the user to draw workflows manually as opposed to ADA systems which generate them automatically based on planning techniques. IDEA uses straightforward search to automatically output the valid processes. The user can select the plan by choosing a ranking method (accuracy, speed, etc.). As opposed to IDEA, the approach in AIDE differs by the way the user and the machine interact: AIDE offers a step by step guidance based on the script planner and the user's decisions. This is suitable for exploratory statistics but it is not suitable for domains where the algorithms run for an extensive period of time.

2.1.3 Re-use past experiences vs. generation from scratch

The Mining Mart project proposed a *case-based reasoning* approach that enables both automatization of preprocessing and reusability of defined preprocessing cases for data mining applications [42]. The best-practice cases of preprocessing chains developed by experienced users are stored and then reused to create new data-mining processes therefore saving time and costs [59]. Moreover Mining Mart includes *cases with self-adapting operators* by using multi-strategy learning [42]. MetaL project developed also a case-based system which combines knowledge and data to advise users which induction algorithm to choose for a particular data-mining task [38]. Recent work in the area [16, 14, 15] tries to combine case based reasoning with ontologies to create a hybrid DM assistant. Their CBR implementation is based on the extension of the classical meta-learning problem from mapping datasets to models, to mapping DM problems to DM cases and on a complementary utility-oriented similarity measure. Also AIDE tries to find similarities between structures to be able to reuse previous results as well as previous decisions [2].

2.1.4 Task decomposition vs. plain plan

Task-oriented user guidance was proposed by Engels and implemented in CITRUS. It guides the users by breaking down the complexity of a typical KDD task and supports him in selecting and using several machine learning techniques. The user guidance module from CITRUS offers assistance in the selection and application of available techniques, interpretation and evaluation of results. It focuses on support and not on automation. AIDE uses hierarchical problem decomposition techniques therefore goals can be decomposed into several subgoals. Problem decomposition and abstraction constitute helpful features for the exploration [52]. CRISP-DM follows a hierarchical process model, having a set of tasks at four levels of abstraction: phase, generic task, specialized task and process instance. The DM process consists of 6 phases, each of them comprise several generic tasks which cover all the possible data mining situations. The specialized tasks describe how the actions from the generic tasks should be accomplished in certain situations. The last level, process instance, represents a record of actions and results of a specific data mining operation.

2.1.5 Design support vs. explanations for result/output

The last version of RapidMiner 5.0 has more user support by introducing the *quick fixes* and the *meta-data propagation* features. Each time the user draws an operator, if it is not connected properly to other operators or some of the input or output fields have incorrect types then the system suggests a set of quick fixes - it gives advice on how the problems could be solved. After loading the data the system extracts and propagates the meta-data such that at any point it can make recommendations. This option can be found in Clementine and in SAS Enterprise Miner (propagation of information) as well. These features represent the support offered to the user during design time. Help buttons are available for each page in IDM. SPSS has more support than other systems for providing *explanations*, the help menu provides extensive information about methods, algorithms, etc. even with examples illustrating the explained feature. Additionally SPSS has coaches that take you step-by-step through the process of interpreting results or deciding which statistical analyses to choose by providing helpful examples - *learning by example* is a very useful feature. SAS Enterprise Miner has integrated debugging and runtime statistics.
REX [24] helps the user in *interpreting intermediate and final results* and also gives useful instructions about statistical concepts. Springex has a primitive 'why' explanation facility which consists of a list of rules that have succeeded together with the conditions that have been asserted. But the knowledge it is unclear - it does not provide explanation of technical terms, is superficial and incomplete. On the contrary Statistical Navigator uses an expert system with help and explanation facilities. Additionally it has extensive reporting capabilities, including a short description of each technique and references to the literature and statistical packages that implement the technique. KENS and LMG provide explanations for concepts but they don't handle interpretation of results or explanation of the reasoning. Contrarily GLIMPSE advises the user on possible interpretations of the output from the statistics package. Moreover it is built on top of a logic-based system shell, APES [31] which offers rule-based explanations - how, why and why not. Another approach is that of IDM which can interpret data by using several statistical measures but it has limited explanation facilities.

2.1.6 Experimental vs. analytical approach

In current systems users can execute workflows, thus the systems provide experimental support (i.e., Clementine, RapidMiner, Weka, etc.) as opposed to SES where the actions are suggested based on the expert knowledge or to MetaL which uses meta-learning. SES and KES are modules developed for existing statistical packages to provide guidance to the users. Therefore the execution is done by those packages. IDM stands between experimental and analytic since it can contain implementation of its own algorithms as well from other data mining systems.

2.1.7 DM experience level

REX can be used not only by expert statisticians but also by naive users. GLIMPSE and KENS are focusing on users with a certain level of knowledge, opposed to LMG which adapts to user expertise. The

[3] Last version of Enterprise Miner is 6.1.You can find a description of the new features here: http://support.sas.com/documentation/onlinedoc/miner/.
[4] RapidI provides RapiMiner: http://rapid-i.com/content/view/181/190/. Last version is 5.0.

inexperienced users can easily feel lost in SPRINGEX and Statistical Navigator since they offer a large amount of knowledge. MULREG is designed for both statisticians and non-statisticians, as well as MLT Consultant and CITRUS which can be used by any domain expert (with any level of DM knowledge). Current DM systems like Weka, Clementine, RapidMiner, etc. have a large set of operators. Even if they have explanations and help facilities it is not trivial for a naive user to solve DM tasks, therefore users need to have experience in using the tool and getting used to the DM domain.

2.1.8 Other features

Visual exploration is a necessity for a system which does data analysis. Therefore most of the existing systems (RapidMiner 5.0, Clementine, DataDesk) as well as exploratory IDAs (AIDE) offer different visualization facilities like icons, interactive graphs, interactive tables, etc. As suggested in [53], *interactivity* is the key to a good EDA, thus it is a desired feature for an IDA as well. Data Desk offers different interactive features like drag and drop, pop-up-menus, hype view menus which provide guidance through potential analyses. Also introduces the concept of extensive user interaction by using links to all graphs and tables.

Hierarchical organization of operators in an ontology, like in IDEA [4], KDDVM [18] and Záková[57], provides several benefits like the inheritance as well as the possibility to introduce abstract operators. An ontology is used as well by [15] to assist novice data miners by storing rules which encode recommendations for DM operators.

Improved navigation plays an important role in user performance in AIDE [1, 3]. Amant has identified navigation functions which lead to improvements in almost any system for statistical data analysis. AIDE includes three types of navigation operations : user actions (go/step, back, forward, any, history, copy/paste, delete, replay), system actions (go/jump, look, refine) and communication actions (overview type, justify,, history, zoom, magnify, filter, landmark, paths).

Ranking of workflows consists an important feature of IDEA - it uses speed and accuracy to provide the user with effective rankings. This facilitates user's decision in the selection of one process that will solve her task. Other work has been done in the field of ranking of classification algorithms based on the accuracy and performance on similar data sets [51] or using different ranking algorithms like average ranks, success rate ratio and significant wins [9].

Preventing serious errors is a feature implemented as well in Clementine and RapidMiner 5.0 by detecting errors based on metadata propagation.

User preferences - IDM stores the history of each user, including user preference and creates profiles for them. Also AIDE [2] incorporates user preferences into the analysis, it records user's decisions and allows her to go backward and forward and to modify her choices. Even MULREG stores and preserves history of the previous models and data structures such that the user can reuse them across different days. KNIME offers a nice feature to the user: a favorite node view where users can manage their favorite nodes, most frequently used and last used.

2.2 Limitations of existing systems

All the systems described in the previous sections are different ways of doing intelligent data analysis, but none of them is a full-IDA system such that it assists the user in doing data analysis.

SES and KES focus on a limited area of DA, thus they work with a limited number of operators. The systems from these two categories were developed during the 80s and are focusing on the needs at that time. Today's needs are much more broad, data analysis includes nowadays a lot of new methods and algorithms for analyzing data. Moreover the size of the data has increased substantially.

Current DA systems as well as SES and KES have less support for automatic workflow design. Even if the user can manually draw the workflow (e.g. RapidMiner, Clementine) for large workflows it can be quite time consuming therefore it would be advisable to generate workflows automatically starting from the features of the input data and the description of the task.

Using most of the systems requires previous knowledge and experience either with the DM operators or with the statistics techniques. Some of the existing systems are developed for naive users (e.g, REX, [16]), to help them solve their tasks easily but then they become too easy for an expert user - they are not helping but restricting her. An IDA should take into consideration all kinds of users and try to integrate facilities for all of them.

Most of the IDAs lack the possibility to reuse similar cases. In DAS users can reopen their previous workflows and try to find out which is the most convenient for the selected task and data but this decision is difficult to be made by a user. It is more a task for the computer - based on algorithms it can figure out which case is most adequate for the current task and data.

Also the explanation support is rather limited and quite superficial. Most of the systems don't have support for interpreting results - this is an advanced feature that would be helpful for the users. Anyway even basic features like description of existing operators are either missing or too scarce.

They either restrict the user in following a set of steps (question-answering systems like REX, KENS) or let him free without any orientation steps (DAS). Restriction is good for naive users - they are usually lost in a jungle of operators or algorithms and it is hard for them to figure out a path to follow. On the other hand expert users prefer to have independence, they know the path to follow but sometimes they have doubts and therefore they need to check with the system if their decisions are correct.

Most of current IDAs don't take user's actions and history into consideration. Users are unique, they use different operators and produce different workflows. It would be helpful to take user's history into consideration and try to learn from it.The user interface could adapt and present to the user only the most used operators, restricting the selection field.

3 Desired features of the future EG-IDA

The IDA should offer explorative guidance to the user - it should help the user discover the Data Mining operators and help him improve the quality of the data mining task and easily solve his task in a shorter time. Besides the features described in Section 2.1, based on the previous work, it is recommended for the assistant to combine the techniques from both mixed-initiative user interfaces [39, 36] and mixed-initiative planning. As argued in [36] the mixed initiative interaction is an essential feature of systems which perform complex tasks as our IDA is doing. Moreover integrating an automated system together with direct manipulation (user's actions) can generate a more natural collaboration between humans and computers. We think that the IDA should at least integrate all the following services to help users increase their productivity:

- **Automatic workflows** - the system automatically generates the

sequence of operators which could solve the user's task. That is helpful for naive users that have little knowledge about the algorithms and methods but also for experts since they can check their knowledge and even find new ways of solving a task.
- **Execution of one or more workflows** - the IDA can find several ways in solving the task so the user should be able to execute alternative workflows at the same time.
- **Step by step execution of a workflow** - the user may want to see how an operator works and what it produces, also she should be able to pause, stop, resume the execution of such a workflow.
- **Ranking of workflows** - for some task the IDA may provide a high number of possible workflows and hence it is difficult for the user to choose one of them. An approach would be to use ranking as suggested in [4] and present to the user the workflows by different criteria : speed, length, accuracy, etc.
- **Reusing workflows** based on their data and goal similarity - it involves case based reasoning - storing all the workflows in a case-base and each time a new workflow has to be generated it will first try to retrieve similar workflows (the similarity will be computed based on the data the user provides and her goal description).
- **Enter goals/hints** and generate proposals of how to reach them
- **Detects errors**, and propose how to repair them
- **Allow the user design his workflows** by using levels of abstraction (in operators, tasks and methods) - the user can use abstractions for operators, methods and tasks and the system can recommend which of the basic instances would better fit the workflow the user has drawn

Moreover it should be compliant to the mixed-initiative interfaces where the user can do anything manually, though :

- At any time she can delegate tasks to the system - the user is not restricted by the system, the user can independently execute her tasks but she can ask the system to solve specific tasks. Also if the task description is not clear enough the system can ask for clarification or can ask questions to the user such that it gets a more specific description.
- In the background the system infers the user's goals and helps to reach them - the system watches over the user's intentions, retrieves all her actions and decisions and based on this information infers what the user wants to achieve and finds possible ways of doing it.
- At any time she can ask the system "what to do next" - at some point the user might not be able to decide what to do next or she has too many decisions that she could make, then the user asks the system to tell her what to do next, therefore the system and the user collaborate and together solve tasks.
- For any decision or step the system can provide explanations - the system offers explanatory information about its decisions and reasoning, tells the user why a specific next step was executed and not another one, thus the user can understand all the system's actions and can learn from them.
- Similar to a spell-checker, the system watches over the correctness and can propose corrections - the system can detect errors and propose fixes, it checks the user's actions and if it finds errors it signals them to the user.

A similar approach is the one described in [47] where the system uses **balanced cooperation** - the modeling task can be done by the user or by a tool of the system. Moreover the user controls the modeling process and guides the learning. There is a synergy between the user and the system, both contribute to model building but the system supports the user, does not take decisions, on the contrary the system is guided by the user. The AIDE system is a mixed initiative system which makes suggestions to the user as well as responds to user guidance about what to do next. The role of mixed-initiative interfaces in intelligence analysis is clearly stated in [7]. Collaboration between the analyst and the system is essential to a better analysis, so the system represents an assistant and not only a tool. As described in [36] the system and the user work as a team, assisting and helping each other but they can also execute some of the tasks independently when required. An important issue is which of the two participants should have the control and when - so when should the system interrupt the user? Mainly the user should have control of the execution and when the system is asked to solve a task it can ask the user more information. Also it would be helpful if the system could provide suggestions to the user when the user is doing nothing or when the user seems lost in the user interface. But the system can help the user intrinsically in any situation without taking control of the execution - for example when the user wants to draw an workflow the system can automatically connect the corresponding nodes or give suggestions for connections.

4 Conclusion

Most of the current IDAs focus on data analysis itself, they offer the users a set of operators to analyze their data. But since the size of the data increased significantly in the last years as well as new types of data have appeared (image, multimedia, etc.) it is hard for users to use the existing tools. There are several attempts to enrich the data analysis systems with guidance but are either too restrictive - offer to the users a set of fixed steps that need to be made, or they provide insufficient guiding. Thus the users have to make decisions based on their own experience and for naive users is not a trivial task.

This paper tried to analyze the IDAs history, identify the problems of existing IDAs and learn from both failure and success. Therefore we introduced a set of metrics based on the existing/proposed implementations of IDAs and looked at their advantages as well as limitations. Moreover we proposed a set of recommended features for designing a new generation of IDAs. The new generation IDA combines techniques from both mixed-initiative interfaces and mixed-initiative planning, therefore it combines automation with user's decisions.

As next steps we intend to implement a new IDA based on the proposed metrics and features. We are convinced that a new IDA is going to fill the gap between analysts/users and todays data analysis systems and also make a faster and better analysis.

ACKNOWLEDGEMENTS

This work is partially supported by the European Community 7^{th} framework program ICT-2007.4.4 under grant number 231519 "e-Lico: An e-Laboratory for Interdisciplinary Collaborative Research in Data Mining and Data-Intensive Science".

REFERENCES

[1] R. Amant, 'Navigation for data analysis systems', *Advances in Intelligent Data Analysis Reasoning about Data*, 101–109, (1997).
[2] R.S. Amant and P.R. Cohen, 'Preliminary system design for an EDA assistant', in *Preliminary Papers of the Fifth International Workshop on Artificial Intelligence and Statistics*, volume 3. Citeseer, (1995).
[3] R.S. Amant and P.R. Cohen, 'Interaction with a mixed-initiative system for exploratory data analysis', *Knowledge-Based Systems*, **10**(5), 265–273, (1998).

[4] Abraham Bernstein, Foster Provost, and Shawndra Hill, 'Towards Intelligent Assistance for a Data Mining Process: An Ontology-based Approach for Cost-sensitive Classification', *IEEE Transactions on Knowledge and Data Engineering*, **17**(4), 503–518, (April 2005).

[5] M.R. Berthold, N. Cebron, F. Dill, T.R. Gabriel, T. Kötter, T. Meinl, P. Ohl, C. Sieb, K. Thiel, and B. Wiswedel, 'KNIME: The Konstanz information miner', *Data Analysis, Machine Learning and Applications*, 319–326, (2008).

[6] R. Bose and V. Sugumaran, 'Application of intelligent agent technology for managerial data analysis and mining', *ACM SIGMIS Database*, **30**(1), 94, (1999).

[7] L.K. Branting, 'The Role of Mixed-Initiative Agent Interfaces in Intelligence Analysis: Extended Abstract', (2005).

[8] P. Brazdil, J. Gama, and B. Henery, 'Characterizing the applicability of classification algorithms using meta-level learning', in *Machine Learning: ECML-94*, pp. 83–102. Springer, (1994).

[9] P. Brazdil and C. Soares, 'A comparison of ranking methods for classification algorithm selection', *Machine Learning: ECML 2000*, 63–75, (2000).

[10] P.B. Brazdil, C. Soares, and J.P. Da Costa, 'Ranking learning algorithms: Using IBL and meta-learning on accuracy and time results', *Machine Learning*, **50**(3), 251–277, (2003).

[11] P.B. Cerrito, *Introduction to Data Mining Using SAS Enterprise Miner*, SAS Publishing, 2007.

[12] B. Chandrasekaran, T.R. Johnson, and J.W. Smith, 'Task-structure analysis for knowledge modeling', *Communications of the ACM*, **35**(9), 124–137, (1992).

[13] P. Chapman, J. Clinton, T. Khabaza, T. Reinartz, and R. Wirth, 'The CRISP-DM process model', *The CRIP–DM Consortium*, **310**, (1999).

[14] M. Charest, S. Delisle, O. Cervantes, and Y. Shen, 'Intelligent Data Mining Assistance via CBR and Ontologies', in *Proceedings of the 17th International Conference on Database and Expert Systems Applications (DEXA'06)*, (2006).

[15] M. Charest, S. Delisle, O. Cervantes, and Y. Shen, 'Bridging the gap between data mining and decision support: A case-based reasoning and ontology approach', *Intelligent Data Analysis*, **12**(2), 211–236, (2008).

[16] M. Charest, S. Delisle, and O. Cervants. Design considerations for a CBR-based intelligent data mining assistant, 2006.

[17] J. Demšar, B. Zupan, G. Leban, and T. Curk, 'Orange: From experimental machine learning to interactive data mining', *Knowledge Discovery in Databases: PKDD 2004*, 537–539, (2004).

[18] C. Diamantini, D. Potena, and E. Storti, 'Kddonto: An ontology for discovery and composition of kdd algorithms', *THIRD GENERATION DATA MINING: TOWARDS SERVICE-ORIENTED*, 13, (2009).

[19] C. Diamantini, D. Potena, and E. Storti, 'Ontology-Driven KDD Process Composition', *Advances in Intelligent Data Analysis VIII*, 285–296, (2009).

[20] W. DuMouchel, 'The structure, design principles, and strategies of Mulreg', *Annals of Mathematics and Artificial Intelligence*, **2**(1), 117–134, (1990).

[21] R. Engels, 'Planning tasks for knowledge discovery in databases; performing task-oriented user-guidance', in *Proceedings of the International Conference on Knowledge Discovery & Data Mining AAAI-Press, Portland, OR*, pp. 170–175, (1996).

[22] R. Engels, G. Lindner, and R. Studer, 'A guided tour through the data mining jungle', in *Proceedings of the 3nd International Conference on Knowledge Discovery in Databases. Newport Beach, CA*, (1997).

[23] Engels, R., *Component-based user guidance in knowledge discovery and data mining*, IOS Press, 1999.

[24] W.A. Gale, 'REX review', in *Artificial intelligence and statistics*, pp. 173–227. Addison-Wesley Longman Publishing Co., Inc., (1986).

[25] W.A. Gale, AT, and T Bell Laboratories., *Artificial intelligence and statistics*, Addison-Wesley Pub. Co., 1986.

[26] J. Gama and P. Brazdil, 'Characterization of classification algorithms', *Progress in Artificial Intelligence*, 189–200, (1995).

[27] K. Gibert, X. Flores, I. Rodríguez-Roda, and M. Sànchez-Marrè, 'Knowledge discovery in environmental data bases using GESCONDA', in *Transactions of the 2nd Biennial Meeting of the International Environmental Modelling and Software Society*, volume 1, pp. 51–56.

[28] K. Gibert, M. Sānchez-Marrč, and I. Rodriguez-Roda, 'GESCONDA: An intelligent data analysis system for knowledge discovery and management in environmental databases', *Environmental Modelling & Software*, **21**(1), 115–120, (2006).

[29] C. Giraud-Carrier, 'The data mining advisor: meta-learning at the service of practitioners', in *Machine Learning and Applications, 2005. Proceedings. Fourth International Conference on*, p. 7, (2005).

[30] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I.H. Witten, 'The WEKA data mining software: An update', *ACM SIGKDD Explorations Newsletter*, **11**(1), 10–18, (2009).

[31] P. Hammond and M. Sergot, 'Augmented PROLOG for Expert Systems', *Logic Based Systems Ltd*, (1984).

[32] D.J. Hand, 'Intelligent data analysis: issues and opportunities', in *Advances in Intelligent Data Analysis. Reasoning about Data: Second International Symposium, IDA-97, London, UK, August 1997. Proceedings*, p. 1. Springer.

[33] DJ Hand, 'Statistical expert systems: design', *The Statistician*, **33**(4), 351–369, (1984).

[34] DJ Hand, 'A statistical knowledge enhancement system', *Journal of the Royal Statistical Society. Series A (General)*, **150**(4), 334–345, (1987).

[35] Hand, D. J., 'Practical experience in developing statistical knowledge enhancement systems', *Annals of Mathematics and Artificial Intelligence*, **2**(1), 197–208, (1990).

[36] M.A. Hearst, 'Mixed-initiative interaction', *IEEE Intelligent systems*, **14**(5), 14–23, (1999).

[37] A. Heckert and JJ Filliben, 'DATAPLOT Reference Manual', *NIST Handbook*, **148**, (2003).

[38] M. Hilario and A. Kalousis, 'Fusion of meta-knowledge and meta-data for case-based model selection', *Principles of Data Mining and Knowledge Discovery*, 180–191.

[39] E. Horvitz, 'Principles of mixed-initiative user interfaces', in *Proceedings of the SIGCHI conference on Human factors in computing systems: the CHI is the limit*, pp. 159–166. ACM, (1999).

[40] R. Ihaka and R. Gentleman, 'R: A language for data analysis and graphics', *Journal of computational and graphical statistics*, **5**(3), 299–314, (1996).

[41] A. Kalousis, J. Gama, and M. Hilario, 'On data and algorithms: Understanding inductive performance.', *Machine Learning Journal, Special Issue on Meta-Learning, , 54(3), 275-312*, **54**(3), 275–312, (2004).

[42] Kietz, J.-U. and Vaduva, A. and Zücker, R., 'Mining mart: combining case-based-reasoning and multi-strategy learning into a framework to reuse KDD-application', in *Proceedings of the fifth International Workshop on Multistrategy Learning (MSL2000). Guimares, Portugal*, volume 311, (2000).

[43] D.M. Levine, M.L. Berenson, and D. Stephan, *Statistics for managers using Microsoft Excel*, Prentice Hall, 1999.

[44] I. MathWorks, 'Matlab', *The MathWorks, Natick, MA*, (2004).

[45] D. Michie, D.J. Spiegelhalter, C.C. Taylor, and J. Campbell, 'Machine learning, neural and statistical classification', (1994).

[46] Ingo Mierswa, Michael Wurst, Ralf Klinkenberg, Martin Scholz, and Timm Euler, 'Yale: Rapid prototyping for complex data mining tasks', in *KDD '06: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 935–940. ACM, (2006).

[47] K. Morik, 'Balanced cooperative modeling', *Machine Learning*, **11**(2), 217–235, (1993).

[48] K. Morik and M. Scholz, 'The MiningMart Approach to Knowledge Discovery in Databases', in *Intelligent Technologies for Information Analysis*, eds., Ning Zhong and Jiming Liu, 47 – 65, Springer, (2004).

[49] J.F.M. Raes, 'Inside two commercially available statistical expert systems', *Statistics and Computing*, **2**(2), 55–62, (1992).

[50] D. Sleeman, M. Rissakis, S. Craw, N. Graner, and S. Sharma, 'Consultant-2: Pre-and post-processing of machine learning applications.', (1995).

[51] C. Soares and P. Brazdil, 'Zoomed ranking: Selection of classification algorithms based on relevant performance information', *Principles of Data Mining and Knowledge Discovery*, 160–181, (2000).

[52] R. St. Amant and P.R. Cohen, 'Intelligent support for exploratory data analysis', *Journal of Computational and Graphical Statistics*, **7**(4), 545–558, (1998).

[53] M. Theus, 'Exploratory data analysis with data desk', *Computational Statistics*, **13**(1), 101–116, (1998).

[54] R. Wirth, C. Shearer, U. Grimmer, T. Reinartz, J. Schlösser, C. Breitner, R. Engels, and G. Lindner, 'Towards process-oriented tool support for knowledge discovery in databases', *Principles of Data Mining and Knowledge Discovery*, 243–253, (1997).

[55] D.E. Wolstenholme and C.M. O'Brien, 'GLIMPSE-a statistical adventure', in *Proceedings of the 10th International Joint Conference on Ar-*

tificial Intelligence (IJCAI 87, Milan, 23-28 August 1987), volume 1, pp. 596–599. Citeseer, (1987).

[56] D.E. Wolstenholme, C.M. O'Brien, and J.A. Nelder, 'GLIMPSE: a knowledge-based front end for statistical analysis', *Knowledge-Based Systems*, **1**(3), 173–178, (1988).

[57] M. Záková, P. Kremen, F. Zelezný, and N. Lavrac, 'Using Ontological Reasoning and Planning for Data Mining Workflow Composition', in *ECML 2008 Workshop on Third Generation Data Mining: Towards Service-oriented Knowledge Discovery*. Citeseer, (2008).

[58] M. Záková, V. Podpecan, F. Zelezný, and N. Lavrac, 'Advancing data mining workflow construction: A framework and cases using the orange toolkit', *ECML PKDD 2009*, 39, (2009).

[59] Zücker, R. and Kietz, J.-U. and Vaduva, A., 'Mining Mart: Metadata-Driven Preprocessing', in *Proceedings of the ECML/PKDD Workshop on Database Support for KDD*. Citeseer, (2001).